A Cross National Investigation of Language Tests: The International English Language Testing System in Europe and Australasia.

Patrick Griffin
Shelley Gillis
Assessment Research Centre
Royal Melbourne Institute of Technology.

Paper presented at the Annual Conference of the Australian Association of Research In Education, University of Newcastle, November 28 to December 3, 1994.

The nature of the International English Language Testing System

The International English Language Testing System (IELTS) was developed as a joint project between the British Council and the International Development Program of the Australian Universities and Colleges (IDP). The system comprises of a series of tests of listening, speaking, reading and writing; developed to measure English language competence and to identify suitable candidates for study in programs conducted in an English language medium. The test battery also enables identification of those overseas students requiring pre-sessional English course training.

Griffin (1988) described the structure, nature and procedures adopted in developing and trialing the test components. In brief, the testing system focused on both productive (writing and speaking) and receptive language skills (listening and reading). The tests were designed to be

clerically scored and the item pool for the tests consisted of a mixture of multiple choice, supply, completion, matching and extended supply. Scales of language proficiency were developed for purposes of score interpretation. The IELTS proficiency scales for each language macro skill had nine levels called "bands". These bands range from 1 to 9 and are described in detail in the specifications of the tests (The British Council, 1988). Band 1 indicates no proficiency and typically represents non users of the English language; whereas band 9 indicates the highest level of language proficiency, roughly equivalent to a native - like proficiency. It was not presumed however that native speakers always scored at the highest levels.

The productive skills tests of spoken language and writing, required the assessors to make judgements using the proficiency band levels, based on detailed criteria and following considerable training. Alternatively, the receptive skills tests of reading and listening required the test scales to be translated into band / scores directly, based on equal intervals on a logit scale (Griffin, 1990; Ingram, 1989).

The reading and writing tests were designed with specific academic populations in mind. A series of specifications for special purpose modules focused on sub populations in academic fields including Science and Technology; Art and Social Sciences; and Life and Medical Sciences. A further set of specifications were also developed to cater for what was described as a non academic, general training population. Table 1 below outlines each tests' composition and schedule.

Table 1 The Test Components and Schedule of Testing in the IELTS Trials.

Population	Code	Component	Items	Time (mins)
General	G1 version 2	Grammar	41	30
*General	G2 version 2	Listening	44	30

General
G3 version 2
Spoken language
n/a
15

*Arts & Social Sciences
M1 version 1
Reading
43
50

Arts & Social Sciences
M1 version 1
Writing
 2
40

*Life & Medical Sciences
M2 version 1
Reading
42
50

Life & Medical Sciences
M2 version 1
Writing
 2
40

*Science & Technology
M3 version 1
Reading
37
50

Science & Technology
M3 version 1
Writing
 2
40

*General Training
M4 version 2
Reading
46
50

General Training
M4 version 2
Writing
 3
40

* Focus of the current report.

All tests were group administered, except the test of Spoken Language which was of an interview format and individually administered. The schedule kept the total testing time at 110 minutes and allowed the full group testing battery to be administered in one sitting. Not all candidates in the trials were required to complete the full battery. The purpose of the trials was to establish the properties of the components and to establish a basis for future reliability and validity studies.

Further, in each test some additional dichotomous closed response questions were asked of the Australasian students. The yes / no responses were used for feedback to the test developers. The questions were:-
SYMBOL 183 \f "Symbol" \s 10 \hDo you feel that this was a fair test of your English?
SYMBOL 183 \f "Symbol" \s 10 \hWas there enough time for you to complete the test?
SYMBOL 183 \f "Symbol" \s 10 \hWas the test too hard?
SYMBOL 183 \f "Symbol" \s 10 \hWas the test too easy?
SYMBOL 183 \f "Symbol" \s 10 \hWere the questions realistic?
SYMBOL 183 \f "Symbol" \s 10 \hWere the instructions clear?
Students were also required to self assess their level of language proficiency on a 9 point scale (i.e., "How well do you think you can understand written English?").

The function of the IELTS system is to average the band scores for all macro skills and develop an overall band score. For most academic courses, the cut off score for admission is an IELTS score at or about Level 6 on the nine point scale (Band 6 refers to a Competent User of the English Language), whilst those for general training programs, typically require an overall band score of at least 5 (Modest User of the English Language). As decisions made on the basis of the IELTS test battery affect potential overseas students' chances of gaining entry to higher education institutions, and can ultimately influence their vocational careers, it is vital that these tests have demonstrated good psychometric properties.

It is in this area of decision making that item response theory (IRT) has considerable advantage over classical test analysis. Reliability estimates supplied by classical analysis do not discriminate over the score range of the test. Whereas with Rasch analysis, IRT estimates

of error are supplied for every score point of the test scale. Hence the academic and training modules should have maximum reliability or minimum measurement error at or about the cut off score of band 6 and band 5, respectively, or the raw scores that convert to those band levels. This was examined in the trials of the test.

Item Response Theory

The item response analysis will be used to examine the fit of the simple logistic Rasch Model to the item level data for each of the tests within samples. The Rasch simple logistic model will be applied to the data to determine the fit of the model to the data. The model is shown in equation 1.

P {xiSYMBOL 110 \f "Symbol"SYMBOL 98 \f "Symbol"SYMBOL 110 \f "Symbol",SYMBOL 100 \f "Symbol"i}= e xij(SYMBOL 98 \f "Symbol"SYMBOL 110 \f "Symbol" - SYMBOL 100 \f "Symbol"i) / [1+e(SYMBOL 98 \f "Symbol"SYMBOL 110 \f "Symbol" - SYMBOL 100 \f "Symbol"i)]

In the model, x represents a score or rating on the item i by person SYMBOL 110 \f "Symbol" . In this simple logistic model, the "x" is restricted to a 1 for correct or a 0 for incorrect. SYMBOL 98 \f "Symbol"SYMBOL 110 \f "Symbol" represents the ability of person SYMBOL 110 \f "Symbol" and SYMBOL 100 \f "Symbol"i represents the difficulty of item i and P represents the probability of obtaining the score of "x" given the person's ability and the difficulty of the item. The Rasch model allows both the item difficulty and the person ability to be directly compared in the same units of measurement, in this case "logits" (Wright and Stone, 1979).

Given the capacity of the Rasch analysis to display the relative levels of item difficulty and person ability, the analysis of the IELTS international pilot study data provided an opportunity to examine person ability and to express this in terms of the item content. The test of fit of the model to the data allows us to examine the dimensionality of the test (Wright and Masters, 1983). Fit indices for all items will be examined across samples to investigate item invariance, potential sample item interaction or bias (Wright and Stone, 1979). Another main advantage of item response theory analysis over traditional test analysis is that if all items can be shown to be measuring the same trait, the candidates ability should be independent of the set of items used to obtain the ability estimate and if the samples of candidates is large, the item difficulty should be independent of the sample of candidates. This latter property is called item invariance and will be tested in this study. It is also possible to estimate the accuracy of all estimates at each score level in the test.

Classical test and item analyses will also be undertaken to identify

such indices as reliability and analyses of variances will be performed to compare the means and variances of tests between samples. This paper is limited to the examination of the international calibration of IELTS receptive test items (reading and listening test components). Cultural differences in response patterns will be identified through data analyses.

The Samples

Trial testing of the IELTS battery was administered by both Australian and British representatives world - wide in 1989. The Australian trials were conducted using a sample of 3,716 non-English speaking students from four Australasian countries:- Indonesia, Thailand, Hong Kong and Australia (referred to as A sample). Alternatively, the British trials were administered on 2,543 non - English speaking Europeans (referred to as the E sample). The countries participating in the British trials were the United Kingdom, Belgium, Cyprus, Egypt, Hungary, Malta, Romania, Hong Kong, Poland, Yugoslavia, France, Sweden and Germany. Table 2 and 3 present the number of candidates assessed on each test in each of the countries from which the E and A samples were drawn, respectively.

Table 2
European Trials:- Sample Sizes for Each Component Test of IELTS and Place of Administration.

	General - Grammar	General - listening	Arts & Social Sciences	Life & Medical Sciences	Science & Technology	General Training	TOTAL
Belgium	0	40	34	7	3	0	84
Cyprus	135	136					

64
0
0
45
380

Egypt
11
12
28
0
0
11
62

Hungary
57
57
33
3
9
10
169

Malta
15
186
98
0
101
0
400

Romania
99
100
0
51
0
0
250

United Kingdom
178
147
121
29
50

105
630

Hong Kong
0
29
18
0
37
0
84

Poland
57
0
14
3
10
0
84

Yugoslavia
25
0
25
0
18
7
75

France
0
0
82
60
20
0
162

Sweden
0
0
14
0
13
54
81

Germany
0
0
0
41
41
0
82

TOTAL
577
707
531
194
302
232
2543

Table 3
Australasian Trials - Sample Sizes for Each Component Test of IELTS and Place of Administration.

General - Grammar
General - listening
Arts & Social Sciences
Life & Medical Sciences
Science & Technology
General Training
TOTAL

Hong Kong
481
465
261
105
113
121
1546

Indonesia
105
106
77
67
73

69
497

Thailand
45
47
8
10
8
21
139

Australia
211
131
271
257
283
381
1534

TOTAL
842
749
617
439
477
592
3716

Results

Each of the receptive tests were analysed using the Quest computer program (Adams & Khoo, 1993) and the summary results of the classical test and item response theory analyses are displayed in Table 4. The receptive tests of the IELTS battery had adequate internal consistency ($0.83 < r2 < 0.94$) given the purpose to which they were to be used, that of recommending admission to Higher education Institutions in the United Kingdom and Australia. The tests also have a wide range of item difficulty and discrimination.

Table 4 also demonstrates that the E and A samples performed similarly on the "Art & Social Sciences" test and the "General Training" test. However, the A sample performed significantly better on the General Tests of "Grammar" and "Listening" than that of the E sample ($t(1417) = -3.51$, $p<.05$; $t(1454) = -4.95$, $p<.05$; respectively).

Alternatively, the E sample produced significantly higher scores on the academic modules "Life & Medical Sciences" and "Science & Technology" ($t(631) = 2.56$, $p<.05$; $t(777) = 10.66$, $p<.05$; respectively), with the greatest effect of cultural differences in test performances being evident for the "Science & Technology" test (effect size = 0.84).

Specific item level data was also analysed for each test. Because of the security of the tests, however, it is not possible to illustrate data using examples of test items. For the purposes of this paper, only data from the "Science & Technology" test were examined for evidence of cultural differences. Table 5 displays the range of item difficulty estimates (logit scores), measurement error (Standard Error of Measurement), the goodness of fit indices (measure of the fit of the model to the item data), the point biserial correlations between item and total score (item discrimination index), and the percentage correct in samples A and E (p-value).

Table 4.Summary test statistics for the IELTS tests of receptive skills, according to place of administration.

Receptive IELTS Tests		
Place of administration		
Sample size		
Number of items		
Reliability co-efficient		
Mean Test Score		
Standard deviation		
Rasch min item logit		
Rasch max item logit		
Point biserial-min		
Point biserial-max		
*Grammar		
Europe		
577		
41		
0.91		
24.62		
7.99		
-2.14		
2.52		
0.11		
0.60		
Australasia		
842		
41		

0.86
25.95
6.24
-3.18
2.75
0.04
0.62

**Listening
Europe
707
44
0.92
21.65
8.85
-2.60
2.89
-0.03

0.58

Australasia
749
44
0.88
23.77
7.46
-2.81
2.91
0.07
0.55

Art & Social Sciences
Europe
531
43
0.85
18.05
6.98
-1.74
2.35
 0.07
0.57

Australasia
617
43
0.90
17.18
8.44
-1.83
2.13
 0.07
0.59

**Life & Medical Sciences
Europe
194
42

0.91
17.87
8.35
-2.62
2.76
 0.01
0.62

Australasia
439
42
0.93
15.85
9.44
-3.06
2.51
 0.19
0.67

**Science & Technology
Europe
302
37
0.90
20.89
7.13
-1.92
3.24
 0.11
0.63

Australasia
477
37
0.91
14.90

7.93
-1.63
3.17
 0.22
0.64

General Training
Europe
232
46
0.84
25.78
6.67
-2.58
2.39
 0.08
0.53

Australasia
592
46
0.84
25.18
6.71

-2.19
2.03
 0.11
0.48

Note:Test score means are significantly different at the 0.05 level, as indicated by two tailed independent t-tests.
Test score means are significantly different at the 0.05 level, as indicated by two tailed independent t-tests and effect sizes (all x's >=-0.20).

Table 5. Science & Technology Item statistics for the A (n=477) and E samples (n=302).

Item name	Logit value	Standard Error	INFIT MNSQ	p.biserial	P value
A					
E					
A					
E					
A					
E					
A					
E					
A					
E					
Item 1					
	-1.33				
	-1.92				
	0.12				
	0.23				
	0.90				
	1.10				
	0.47				
	0.26				
	75.5				
	92.4				
Item 2					
	0.32				
	0.33				
	0.11				
	0.14				
	0.87				
	0.97				
	0.61				
	0.46				
	46.0				
	65.2				

Item 3
0.45
0.47
0.12
0.14
0.87
1.00
0.60
0.46
45.9
63.8

Item 4
-0.25
0.34

0.11
0.14
1.15
1.19
0.39
0.32
58.1
65.3

Item 6
-1.63
-1.50
0.13
0.21
 0.95
0.86
0.39
0.43
80.2
89.6

Item 7
-0.98
-1.22
0.12
0.19
0.91
0.83
0.45
0.48
70.5
86.9

Item 8
1.05
-0.78
0.12
0.17
1.27
1.01
0.30
0.37
33.5
82.6

Item 9
-1.28
-0.95
0.12
0.18
0.94
0.85
0.39
0.48
75.6
84.1

Item 10
-0.97
-0.66
0.12
0.17

0.96
0.80
0.44
0.57
70.6
80.7

Item 11
-0.70
-0.48
0.11
0.16
0.92
0.92
0.47
0.49
66.1
78.4

Item 13
-1.34
-1.78
0.13
0.23
0.86
0.89
0.47
0.38
76.6
91.6

Item 14
0.12
0.22
0.11
0.14
0.96
1.05
0.49
0.40
51.0
67.7

Item 15
-0.86
-0.96
0.12
0.18
0.83
0.86
0.54
0.50
69.7
84.2

Item 16
0.35
0.43
0.11
0.14
0.89
0.97

0.56
0.50
46.7
63.8

Item 18
-0.80
-1.01
0.12
0.18
0.98
1.14
0.46
0.27
68.4
85.0

Item 19
-1.31
-1.75
0.13
0.23
0.92
0.89
0.45
0.39
76.6
91.6

Item 20
0.37
0.44
0.12
0.14
0.92
1.06
0.57
0.44
46.4
63.7

Item 21
-0.31
0.08
0.12
0.15
1.18
1.11
0.35
0.36
60.7
70.0

Item 22

-0.77
-0.49
0.12
0.17
1.15
1.05
0.33
0.42

69.0
78.8

Item 23
-0.56
-0.49
0.12
0.16
1.03
0.99
0.42
0.42
63.8
78.1

Item 24
0.92
0.42
0.12
0.14
1.04
0.97
0.46
0.49
36.9
63.7

Item 25
0.91
1.12
0.13
0.14
1.39
1.19
0.22
0.34
37.7
50.3

Item 26

-0.01
0.89
0.13
0.15
1.54
1.45
0.06
0.11
57.6
57.3

Item 27
3.11
3.24
0.20
0.19
1.08
1.17
0.33
0.28
11.7
18.1

Item 29
0.87
0.63
0.14
0.15
0.91
0.82
0.59
0.62
41.3
61.9

Item 30
0.38
0.77
0.13
0.15
1.05
0.96
0.46
0.52
50.2
59.6

Item 31

0.74
1.11
0.14
0.15
1.05
1.08
0.47
0.43
44.5
53.7

Item 32
0.35
0.17
0.14
0.17
0.84
0.83
0.62
0.59
52.9
71.4

Item 33
1.21
1.60
0.15
0.15
1.19
1.15
0.36
0.34
36.4
45.3

Item 34

0.19
0.23
0.15
0.17
1.00
1.11
0.51
0.38
54.4
70.2

Item 35

0.89
0.88
0.16
0.16
0.87
0.87
0.61
0.58
43.0
59.5

Item 36
0.49
0.83
0.15
0.16
0.82
0.82
0.64
0.63
50.8
60.7

Item 37
0.36
-0.20
0.15
0.19
0.89
0.94
0.58
0.51
53.6
78.5

The test width of the "Science & Technology" module is relatively large . Apart from Item 26 (pt biserials < 0.12 for both samples), all items had adequate point biserials regardless of sample and the Rasch model was shown to fit the data for all items with exception of Item 26 (for both samples) and Item 25 (for the A sample). This enables a conclusion that the test was uni dimensional and that the items were acting as a cohesive set measuring a single dimension of language independent of country of origin or of first language spoken. This, according to Wright and Masters (1982) is evidence of construct validity. Item 26 data was further analysed to determine why it was not consistent with the underlying trait and why it was a poor discriminator of ability levels for both Europeans and Australasians.

The item required students to produce a dichotomous response to a given passage of writing. This response however, required prior knowledge of the subject matter. It was recommended that the item be eliminated from the final version of the test. Prior knowledge could also be used to answer Item 25. Hence, it should also be excluded from the final version of the test.

The test width is more clearly demonstrated graphically, where the distribution of scores of the students are plotted relative to the distribution of the difficulty levels of the items on the test (refer to Figure 1 for the A sample distribution and Figure 2 for the E sample distribution). The distributions maps have three scales. The first is the raw score of the students, the second the latent trait logit scale (with standard error of measurements reported), and the third is the band scale for interpretation of the IELTS. These scales all indicate the same measure. Scores can be converted from raw scores to logits of ability or difficulty, and to the descriptive scale of the IELTS. The band score of 6 on the "Science and Technology" module is established as the cut off score for admission to academic courses and the standard error at about this level is lower than other levels indicating that the most accurate decisions can be made at this score level as required.

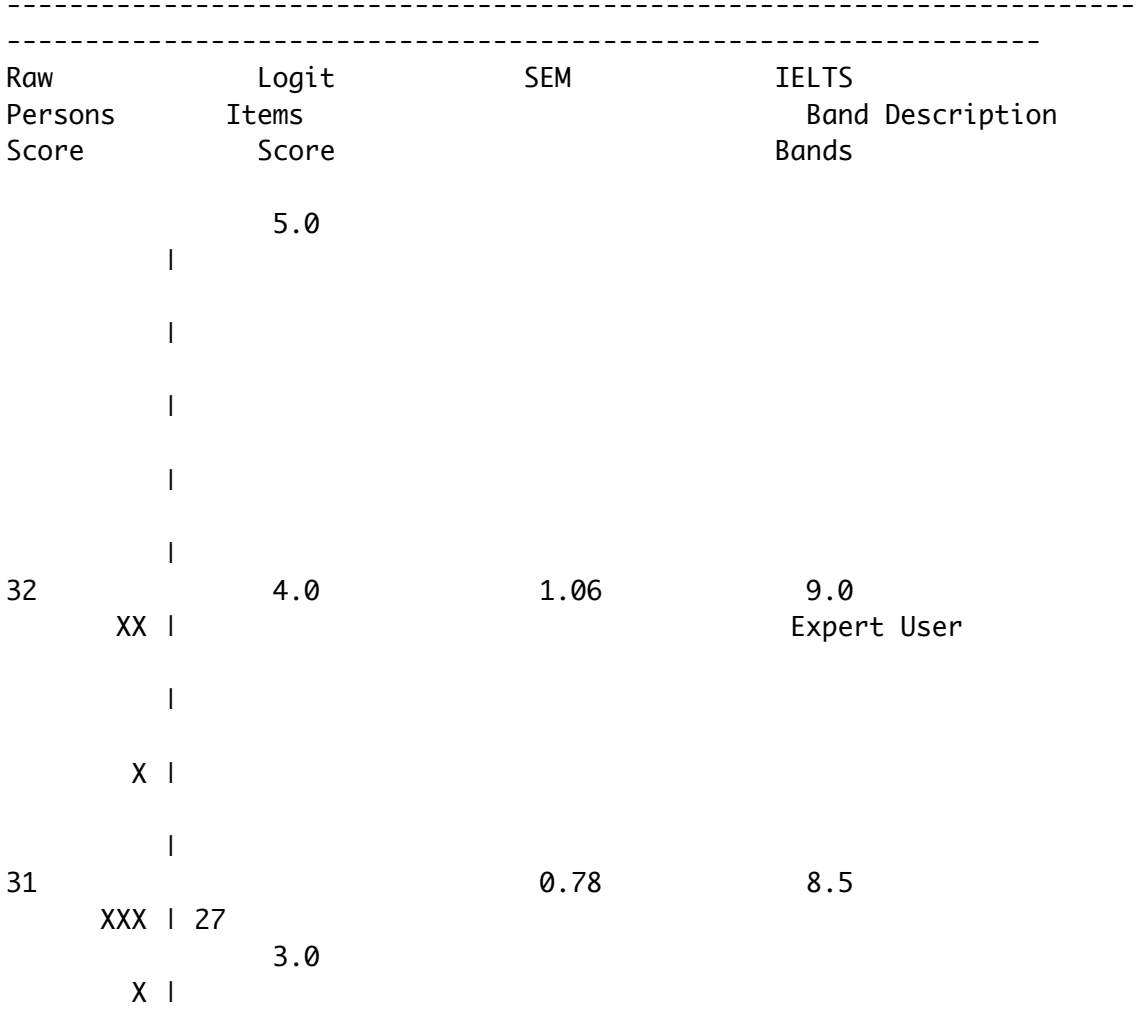
An examination of both distributions clearly demonstrates that the test difficulty range matched the ability range of the students in both samples A and E. However, the E sample scored significantly higher than the A sample (E SYMBOL 109 \f "Symbol"=20.89, SYMBOL 115 \f "Symbol"=7.13; A SYMBOL 109 \f "Symbol"=14.90, SYMBOL 115 \f "Symbol"=7.93). Closer examination of the item statistics in Table 5, reveals possible evidence of an interaction between sample location and item difficulty. For instance, apart from Item 27 and Item 33, over half of the E sample provided correct responses to all items. Alternatively, greater than 50% of the A sample answered 7 out of the 33 items incorrectly (i.e., items 2, 3, 8, 23, 24, 27 & 33).

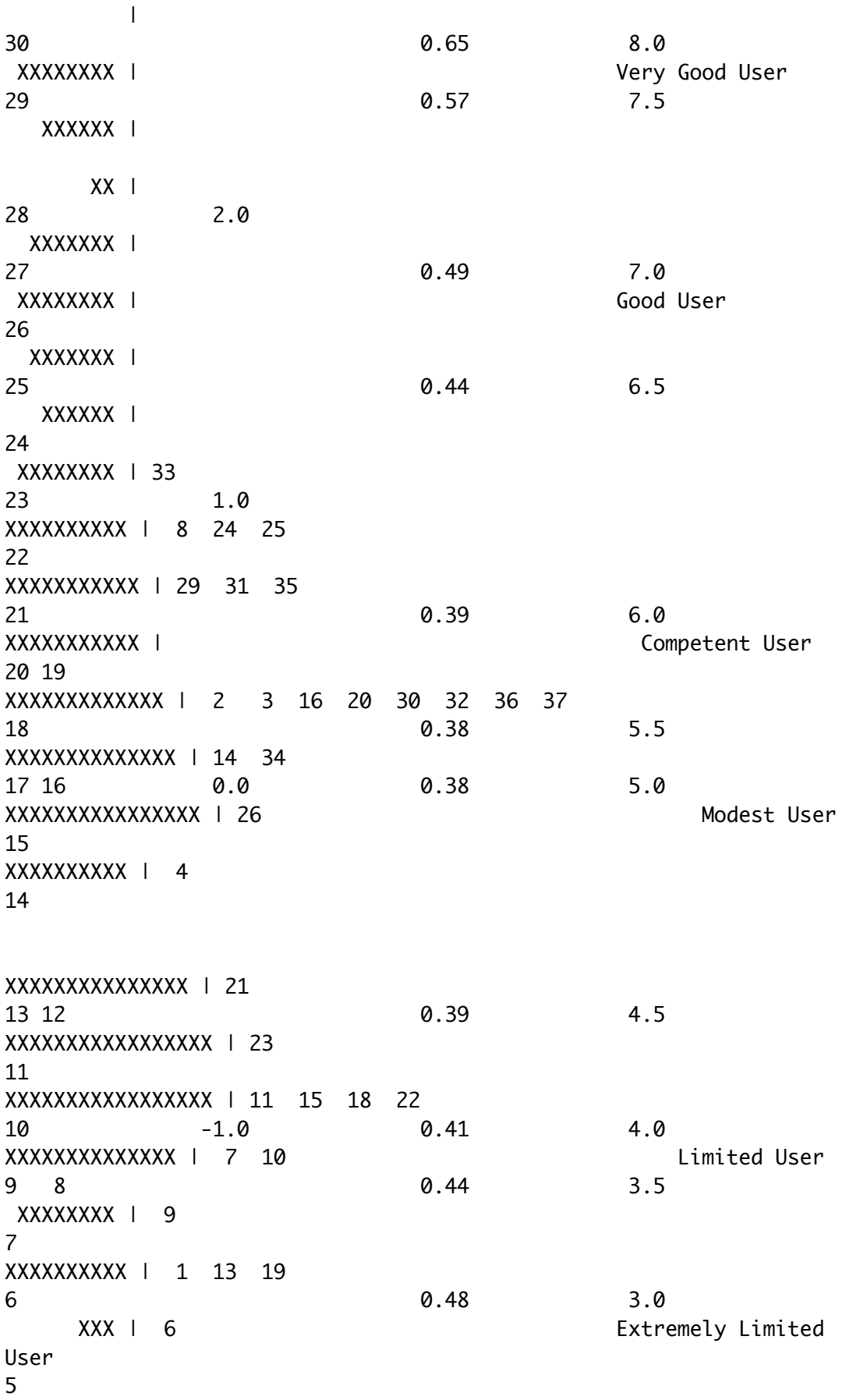
While the test width appears to be appropriate for the student group, there is a potential difficulty associated with the spread of items and the use to which the test is put in some institutions. Many universities set higher band levels for entry into post graduate degrees. For instance, in the United Kingdom, some universities require band levels of 7 or 7.5 even for undergraduate degrees.

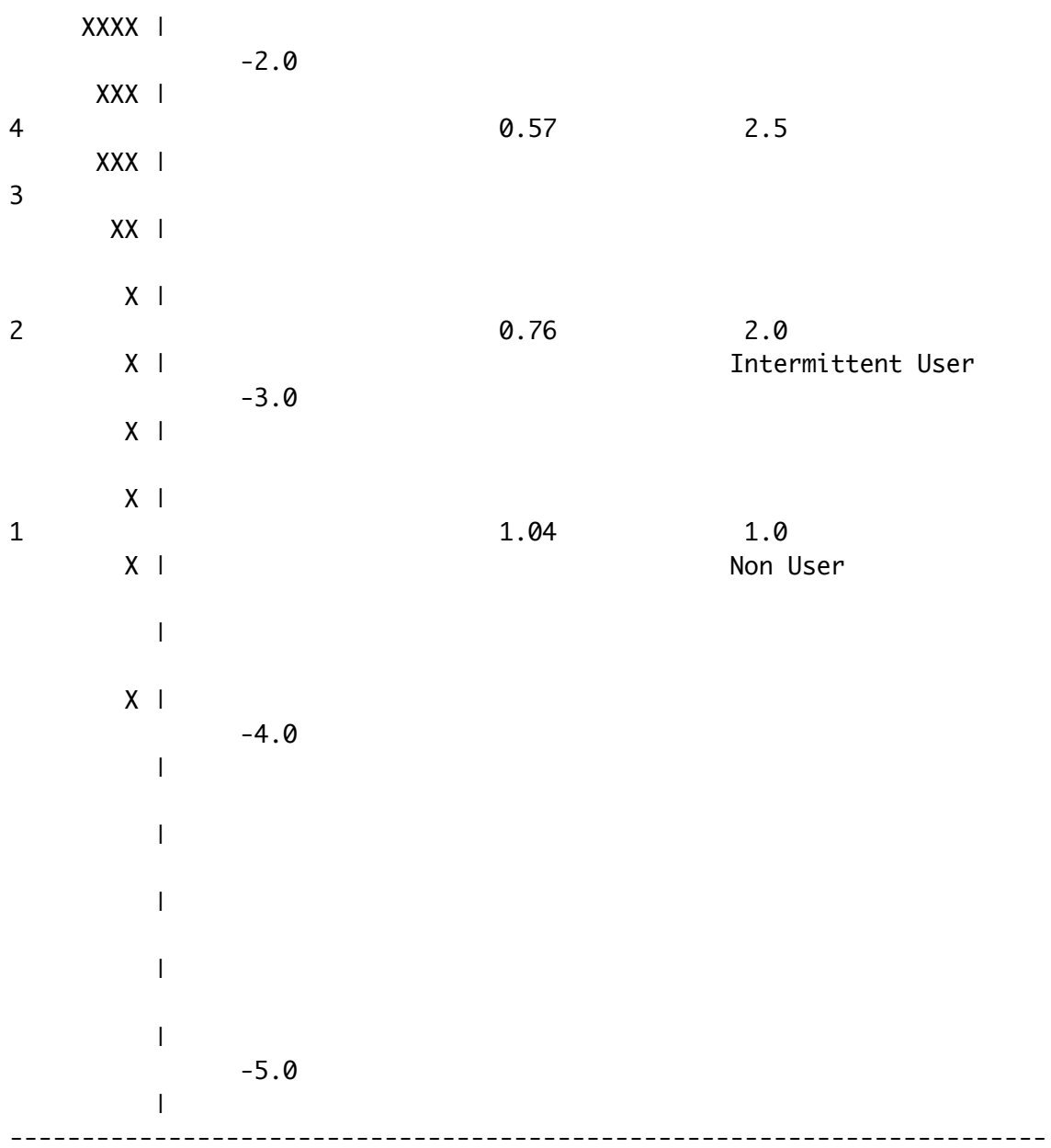
While these levels select out students with very high proficiency levels, the variable maps, shown in figures 1 and 2, show a lack of items at the these band levels. While it is possible to translate the raw scores to bands at these levels for decision making, it is clear that the amount of information available for decision making is limited. This is illustrated by the gap in the spread of items over these bands. In fact the test does not have any items at the upper band levels. So while it is possible to discriminate at the band 5.5

to 6.5 with some confidence in the accuracy of the test, it is not possible to make the same confident discrimination at the higher levels. Additional items should have been written to cover these gaps in the test spread of items.

This may have been difficult however because of the approach to the development of reading comprehension items. Single items are rarely developed. Instead, sets of items are matched to passages, to documents or to other stimulus materials. The sets may consist of a series of matching items, multiple choice or supply items. Variability in item difficulty was not controlled, nor generally directed towards band levels on the test. Instructions were given to items writers to match band levels but this proved difficult for the writers. It may be that the test needs to be analysed using sets of items rather than individual items and this is the subject of later research, based on this initial analysis.



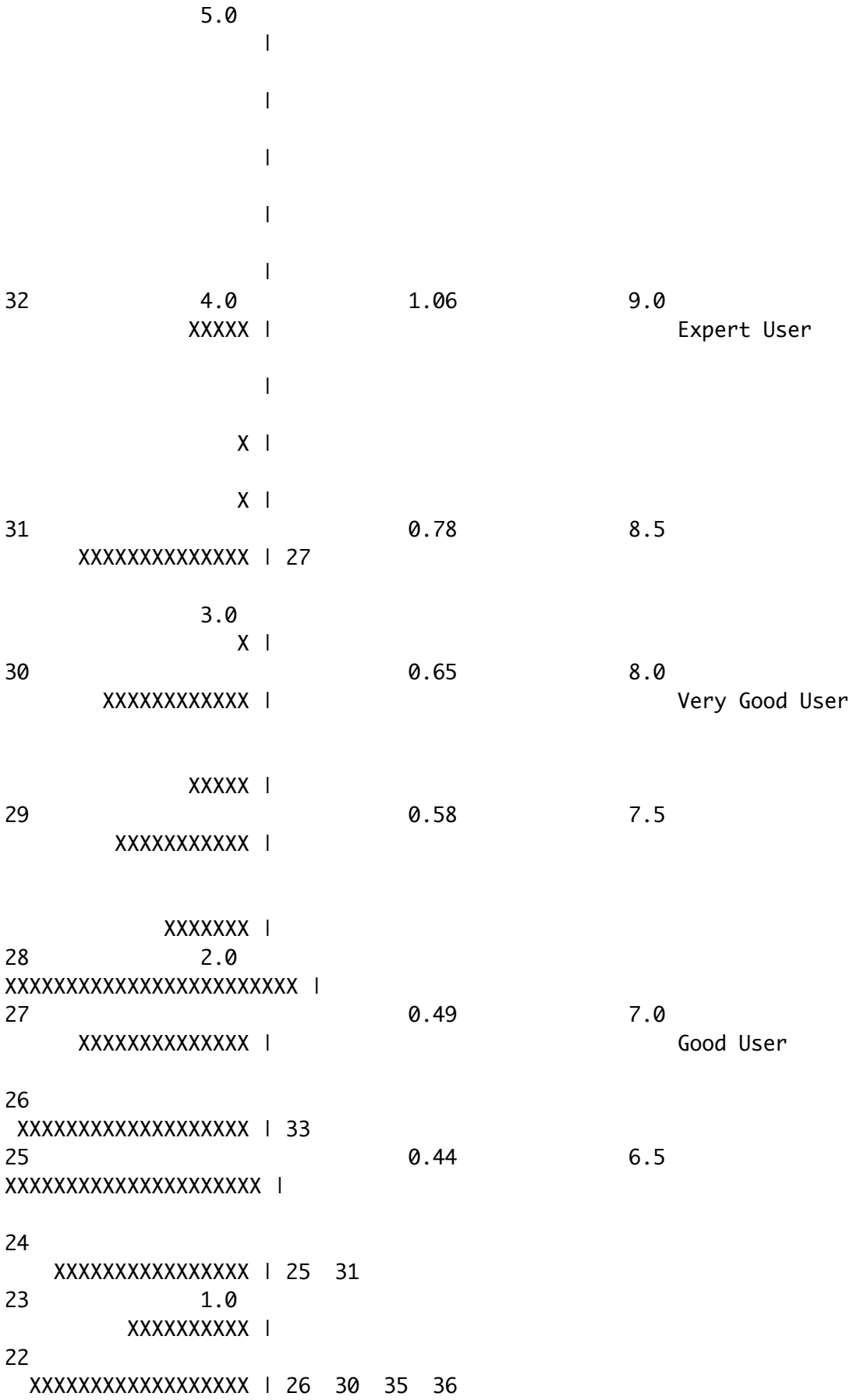


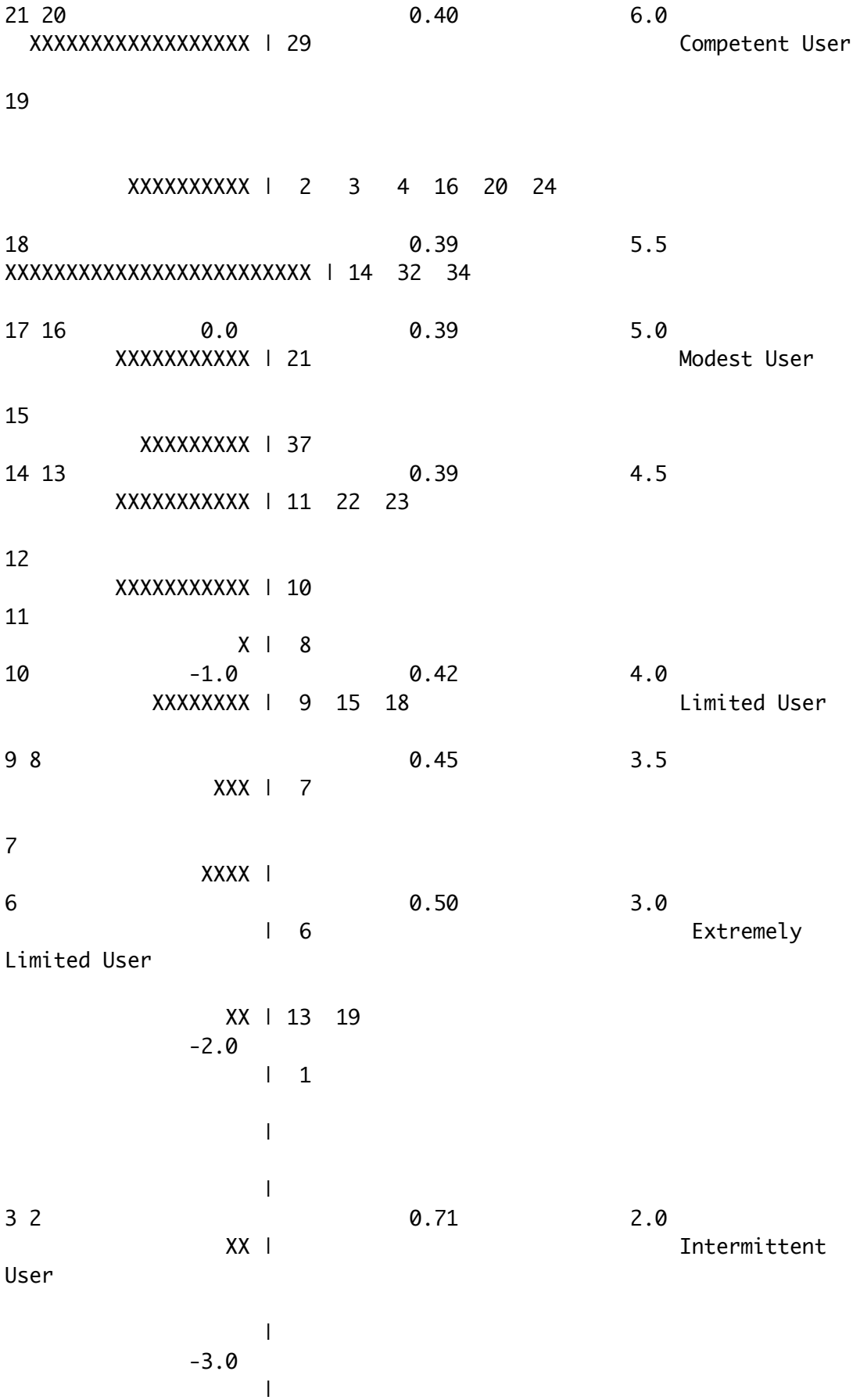


Each X represents 2 students

Figure 1.Science & Technology Module: Conversion from Raw scores to Band levels:- Australasian Sample

Raw Score	Logit Persons Score	Items	SEM	IELTS Bands	Band Description
4	2.5		0.57	2.5	
3					
2	2.0		0.76	2.0	Intermittent User
1	1.0		1.04	1.0	Non User





```
                    |
                    |
                    |
                    |
              -4.0
                    |
                    |
                    |
                    |
                    |
              -5.0
                    |
---------------------------------------------------------------------
---------------------------------------------------------------------
-
  Each X represents    1 students
=====================================================================
=====================================================================
=
```

Figure 2.Science & Technology Module: Conversion from Raw scores to Band levels :- European Sample

To examine item invariance across cultures, linear regression analyses were performed on the both the item calibrations and the infit mean square estimates. The scatter plots, with 95% confidence bands are displayed in Figures 3 & 4 respectively.

LINK SPSSCHRT C:\\SPSSWIN\\IELTS\\M3V1LOG.CHT Contents * mergeformat \p \a

Figure 3.Item Calibrations as a function of cultural background - Science & Technology Module.
Linear Regression Scatterplots.

LINK SPSSCHRT C:\\SPSSWIN\\IELTS\\M3V1MSQ.CHT Contents * mergeformat \p \a

Figure 4.The fit of the item to the model as a function of cultural background - Science & Technology Module.
Linear Regression Scatterplots.

Item 8 appears to have interacted with cultural background. In regards to the E sample, the item was relatively easy to perform and could be completed correctly by a "Modest User" (refer to Figure 2). In contrast, the item was rather difficult for the A sample (34% with correct responses in comparison to 83% of the E sample, refer to Table 4). In fact, only "Good Users" in the A sample typically produced a correct response. A closer content examination of the item reveals that it is one of a set of matching items allowing multiple use of matching elements. The mis fit may be related to the interdependence among the items. Items 9 and 10 depend on the correct answer being given for item 8. Item 8 asks the student to recognise and match a general recommendation given in the text. Items 9 and 10 then ask the student to recognise reasons in the text given for that general recommendation. Moreover there is only one alternative in the matching set that is written in the form of a recommendation. It is therefore possible to answer the item 8 without reading the passage! It points to test wiseness being a possible explanation for the E sample response pattern as distinct to the A sample response pattern, for this type of item. The solution may be in the preparation for the test rather than in the redevelopment of the students. E sample students may have been more familiar with the IELTS style of testing which differs from the alternative test used in Australia and Asia prior to the introduction of the IELTS:- The College Board's Test Of English as a Foreign Language (TOEFL) relies much more on multiple choice tests items.

Item 8 also lacked invariance of model fit between samples (refer to Figure 4). The slight overfit (high positive mean square residual) (Wright and Masters, 1982), suggests that a sharp distinction could be made at a specific ability level between those who were correct and those who were not. Only item 8 demonstrated item invariance with respect to fit and difficulty using the 95% confidence interval about the regression of the E sample estimates on the A sample estimated. Given that the test contained 33 items, the results indicate a very stable test of English reading proficiency.

Similar analyses for the listening tests were carried out and these analyses have been repeated for all modules.

Student satisfaction with the test was assessed using the series of feedback questions previously discussed. The results of these items together with a student self assessment on a 9 point scale of reading proficiency are provided in Table 6.

Table 6: Science and Technology Test of Reading and Writing: General Properties and Student Feedback (N = 477)

Variable P-Value Std Dev r.pbi
Do you feel that this was a fair test of your English? 0.72 .449 +.204
Was there enough time for you to complete the test? 0.27 .442 +.465
Was the test too hard? 0.52 .500 -.560
Was the test too easy? 0.07 .247 +.275
Were the questions realistic? 0.83 .373 +.128
Were the instructions clear? 0.84 .367 +.221

Variable Mean Std Dev r.pbi
Self Rating 1 - 9. 4.897 1.333 +.284

The student evaluation of the tests, performed by the Australasian sample, supported the overall success of the test as a measure of the English proficiency for admission to Institutions of higher education. Only seven percent of the Australasian's sampled, thought that the test was too easy. In summary, high scorers tended to agree that the test was fair, that there was enough time to complete for completion, that the test was not too hard, that it contained realistic items, and had clear instructions. Low scorers however tended to disagree and gave the test a negative evaluation. There was not a high correlation between the self assessment and the test scores, indicating that self assessment was not an adequate alternative method. In general the students tended to cluster their estimates around the middle of the scale perhaps indicating that they were unsure of their proficiency level. The mean rating was 4.9 on a 9 - point scale.

Conclusion:

The Science and Technology Reading test shows many characteristics of a stable and reliable test, with construct validity. Item invariance can be shown to exist over international samples. The test had appropriate width for the sample of students and had demonstrated high classical reliability with low errors of measurement at and near the cut point recommended by the developers for institutions to admit academic overseas students. The test also provided evidence of the fit of the Rasch simple logistic model to the data, providing support for the single dimension nature of the test required by the test developers. Finally, the students showed a high level of satisfaction with the test.

The development of the IELTS was a large undertaking, involving numerous item writers in several countries. Pilot studies were conducted in seventeen countries and a total sample size of more than 6000 students were involved in the pilot study. Given these parameters, the pilot study simulated the range of cultures and admission groups for which it was designed. The evidence in this study indicates that the test has the properties that will enable it to fulfil its role of making admissions to higher education institutions possible and based on adequate data which has high levels of accuracy and reliability.

References:

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Adams, R.J., and Khoo, S.T. (1993). Quest:- The Interactive Test Analysis System. Australian Council for Educational Research: Melbourne.

Griffin, P., (1988). English Language Testing Service (ELTS) Revision: Procedures and Products. Paper presented at the Annual Conference of the Applied Linguistics Association of Australia: Launceston.

Griffin, P., (1990). Characteristics of the Test Components of the IELTS Battery: Australian Trial Data. Paper presented at the Regional English Language Centre, Annual Seminar: Singapore.

Ingram, D.E., (1990). The International English Language Testing System (IELTS): It's Nature and Development. Paper presented at the Regional Language Seminar Language Testing and Programme Evaluation: Singapore.

Ingram, D.E., and Wiley, E. (1979, revised 1985). The Australian Second Language Proficiency Ratings. Griffith University, Mimeograph.

Rasch, G. (1960, revised 1980). Probabilistic models for some intelligence and attainment tests: University of Chicago Press: Copenhagen: Danmarks Paedegogiske Institute and Chicago.

Skeehan, P. (1988) Peter Skehan on Testing. Part I. Language Teaching, 21(4), 211-221.

Skeehan, P. (1989) Peter Skehan on Language Testing. Part II. Language Teaching, 22(1), 1-13.

The British Council (1988). International English Language Testing System. An introduction to IELTS. The British Council: Cambridge.

Wright, B., and Masters, G. (1982). Rating Scale Analysis. MESA Press: Chicago.

Wright B., and Stone, M. (1979). Best Test Design. MESA Press: Chicago.